基于长短期记忆网络的中文命名实体识别

管浩宇[1]

[1]（北京化工大学信息学院 北京 10010）

摘要：

命名实体识别是识别文本并理解的重要步骤，经过长久发展，使用深度学习方法研究和改进命名实体识别过程已成为常见方法。本项目使用基于长短期记忆网络（LSTM）模型的方法，实现了一个中文命名实体识别系统，并实现了较高准确率。

关键词：中文命名实体识别　深度学习　长短期记忆网络

分类号：TP181

Chinese Named Entity Recognition Based on Long Short Term Memory Networks

Guan Haoyu[1]

[1](College of Information Science and Technology, Beijing University of Chemical Technology, Beijing 10010, China)

Abstract：

Named entity recognition is an important step in recognizing text and understanding it, and after a long development, it has become a common approach to study and improve the named entity recognition process using deep learning methods. This project implements a Chinese named entity recognition system using an LSTM model-based approach and achieves a high accuracy rate.

Keywords： Chinese Named Entity Recognition Deep Learning LSTM

一、项目目标

命名实体识别是指识别出文本中具有特定意义的命名实体并将其分类为预先定义的实体类型，如人名、地名、机构名、时间、货币等。在大数据时代，如何精准并高效地从海量无结构或半结构数据中获取到关键信息，这是自然语言处理任务的重要基础。命名实体通常包含丰富的语义，与数据中的关键信息有着密切的联系，NER任务可以用于解决互联网文本数据的爆炸式信息过载问题，能有效获取到关键信息，并广泛应用于关系抽取、机器翻译以及知识图谱构建等领域。本项目鉴于命名实体识别任务的重要性，结合学习内容和自学资料，拟实现一个中文的命名实体自动识别系统，使用基于LSTM模型的方法，实现命名实体识别的功能，并结合窗口展示功能，完成系统。在获取的数据集上进行实验，识别文字中的专有名词：人名、地名、组织机构名，提取文字中丰富的语义信息，可以在应用场景中达到良好

的效果。

二、国内外相关工作

自1991年第7届IEEE人工智能应用会议上Rau发表了一篇“从文本中抽取公司名称”的论文，提出了一种从文本中提取公司名的方法，经过30年NER的发展历程，主流的NER方法可以分为3类：基于规则和词典的方法、基于统计机器学习的方法和基于深度学习的方法。这3类方法根据处理特点又细分为若干种不同的子方法[1]。

2.1 基于规则和词典的 NER 方法

早期的NER方法主要运用由语言学专家根据语言知识特性手工构造的规则模板，通过匹配的方式实现命名实体的识别。1995年，Krupkal[2]提出了一个用于英文NER的SRA系统，包括NameTag和 HASTEN两个子系统。面向中文的NER起步较晚，1997年，张小衡等[3]根据机构名称的结构规律和形态标记等特点进一步总结规则，从600多万的三地语料库中识别高校名称实体，正确率达到了97.3%。2002年，王宁等[4]从专业名词识别的角度，针对金融新闻文本，充分考虑金融领域的特征，利用规则的方法专门针对公司名的识别问题进行了研究，总结了公司名的结构特征以及上下文信息，归纳形成知识库，并采取两次扫描的策略进行识别。

基于词典和规则的实体识别方法使用简单，结果准确率较高，特别是对于数字和时间日期实体利用规则匹配的方式能获得较好的识别效果。但是词典和规则库的建立需要花费大量时间和人力;不同的实体类型需要定制相应的规则，移植性差。为了解决上述问题，一些专家学者研究了统计机器学习的实体识别方法[5]。

2.2 基于统计机器学习的NER方法

（1）有监督学习

有监督学习的NER方法是将NER任务转换成分类问题，通过机器学习方法将已标记的语料构造为特征向量，以此建立分类模型来识别实体。采用有监督机器学习的分类模型包括:HMM(hidden Markov models)、MEM(maximum entropy mo-dels)、SVM (support vector machines)和CRF(condi-tional random fields)等模型。

基于HMM的NER方法利用维特比算法将可能的目标序列分配给每个单词序列，能够捕捉现象的局部性，进而提高了实体识别性能。Bikel等[6]基于大小写、数字符号、句子首词等特征，利用HMM来计算某一单词为某一实体类型的概率。Zhou等[7]提出一种基于HMM的组块

标记器的 NER 方法，在 Bikel 的基础上扩充了内部语义特征、内部地名词典特征以及外部上下文特征，对 HMM 的传统公式做了改进，以便能融合更多的上下文信息来确定当前预测类型。对于中文 NER，俞鸿魁等[8]提出一种基于层叠 HMM 的中文 NER 模型，该模型由三级 HMM 构成。

基于 MEM 的 NER 方法的主要思想是在已知部分知识的前提下选择熵最大的概率分布，从而来确定某一实体的类型，MEM 能够较好地融合多种特征信息进行分类。Borthwick 等[9]最早将 MEM 用于英文 NER 任务，综合考虑了首字母大小写、句子的结尾信息以及文本是否为标题等多种特征信息。对于中文 NER，张明杰等[10]提出一种融合多特征的 MEM 中文 NER 模型，该模型能集成局部与全局多种特征，将规则和机器学习的方法相结合，分别构建了局部特征模板和全局特征模板，同时引入启发式知识解决效率和空间问题。

SVM 是定义为特征空间上的间隔最大的线性分类器。首先通过高维特征空间的转化使分类问题转换成线性可分问题，然后基于结构风险最小理论构建最优分割超平面，使得分类器得到全局最优化。该模型在 NER 任务上被广泛使用，Isozaki 等[11]提出了一种基于 SVM 的特征选择方法以及有效的训练方法，能增加系统训练的速度。陈霄等[12]针对中文组织机构名的识别任务，提出了一种基于 SVM 的分布递增式学习的方法，利用主动学习的策略对训练样本进行选择，逐步增加分类器训练样本的规模，进一步提高分类器的识别精度。

CRF 模型统计了全局概率，不仅在局部进行归一化，且考虑了数据在全局的分布情况。CRF 具有表达长距离依赖性和交叠性的优势，能有效融入上下文信息以及领域知识，可以解决标注偏置问题。即使 CRF 具有时间复杂度高导致的训练难度大等问题，但仍十分广泛地被用于 NER。McCallum 等[13]提出了一种基于 CRF 的特征归纳的 NER 方法，与传统方法相比，自动归纳特征既提高了准确性，又显著减少了特征数量。燕杨等[14]针对中文电子病历的 NER 问题，提出一种层叠 CRF，该模型在第二层中使用包含实体和词性等特征的特征集，对疾病名称和临床症状两类命名实体进行识别。

（2）无监督学习

为了解决跨域和跨语言标注文本的不足，学者们提出了 NER 的无监督学习技术。无监督学习是不需要使用标注数据的算法，该方法使用未标注的数据来做出决策。无监督学习旨在考虑数据的结构和分布特征，从而发现更多关于数据的学习。比如，2016 年，Han 等[15]提出一个基于聚类主动学习的生物医学 NER 系统，该聚类方法通过使用底层分类器在文档中查找候选命名实体来进行聚类，因而更能反映命名实体的分布。

2.3 基于深度学习的 NER 方法

基于深度学习的 NER 方法一般流程如图所示，共分为 4 步[16]：(1)Sequence，预处理后

的输入序列。(2)Word embedding ，将输入序列转换成固定长度的向量表示。(3)Context encoder ，将词嵌入进行语义编码。(4)Tag decoder，进一步进行标签解码。

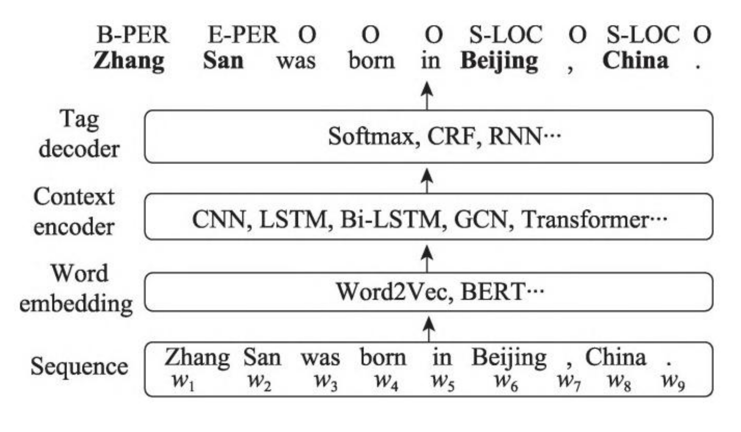


表 1 基于深度学习的 NER 一般流程[1]

2011 年，Collo-bert 等[17]提出了一种基于 CNN 的 NLP 模型，能处理包含 NER 等多种任务。2021 年，Kong 等[18]提出一种融合多层次 CNN 和注意力机制的中文临床 NER 方法。该方法既能捕捉短距离和长距离的上下文信息，且注意力机制还能获取全局上下文信息，进一步解决了 LSTM 在句子较长时无法捕捉全局信息的问题。近年来，GCN (graph convolutional network)和 GGNN(gated graph neural network)在 NER 任务中得到广泛的关注。Cetoli 等[19]率先在 NER 任务中使用图 GCN 来解决实体识别问题，在传统的 Bi-LSTM-CRF 模型的 Bi-LSTM 层和 CRF 层中间额外添加一层 GCN 层，利用句子的句法依存关系构图，通过 GCN 将节点信息传递给最近的节点，通过将 N 层图堆叠在一起，可以传播最多相距 N 跳的节点特征。基于 Transformer 方法典型代表是 BERT 类的预训练模型。Souza 等[20]在 NER 任务上提出一种 BERT-CRF 模型，将 BERT 的传输能力与 CRF 的结构化预测相结合。Yang 等[21]提出了一种分层的 Transformer 模型，应用于嵌套的 NER。基于深度学习的 NER 方法中，Transformer 方法代表模型为 BERT，注意力层能够利用句子中标签顺序及出现规律进行标签预测，解决了神经网络仅根据单个字向量做标签预测的问题，其优越性在于它能够给各单元分配一个相关性权重，能够显性地分辨哪个词对实体正确表示最为关键，有效地剔除噪音，进一步提升模型效果。采用掩码语言模型对双向的 Transformer 进行预训练，以生成深层的双向语言表征，但需要大量 GPU 和训练数据。

三、实现系统（或模块）的核心思想和算法描述

在文献调研过程中，我们关注到有关 LSTM 模型的命名实体识别相关工作，该种模型在

命名实体识别领域具有十分广泛的应用，英文 NER 领域发展较快，因项目目标定为拟实现中文命名实体识别工作，所以在关注中文的 NER 发展情况过程中，发现了 LSTM 在中文 NER 中的逐步应用，LSTM 是 RNN 的一种变体，其核心概念在于细胞状态以及“门”结构。细胞状态相当于信息传输的路径，让信息能在序列连中传递下去；“门”结构在训练过程中会去学习该保存或遗忘哪些信息。实现系统采用了双向 LSTM 模型进行命名实体识别，使用 python 编程语言，采用 pytorch 机器学习库实现。模型的主要框架由一个嵌入层、一个 LSTM 层和一个全连接层（分类）组成，代码如下：

```
class BiLstm(nn.Module):
    def __init__(self, corpus_num, embedding_num, hidden_num, class_num, 
bi=True):
        super().__init__()
        self.embedding = nn.Embedding(corpus_num, embedding_num)
        self.lstm = nn.LSTM(embedding_num, hidden_num, batch_first=True, 
bidirectional=bi)
        if bi:
            self.fc = nn.Linear(hidden_num * 2, class_num)
        else:
            self.fc = nn.Linear(hidden_num, class_num)
    def forward(self, batch_data):
        embedding = self.embedding(batch_data)
        out, _ = self.lstm(embedding)
        return self.fc(out)
```

该模型通过交叉熵损失函数在训练集上计算误差，再进行反向传播，使用 Adam 优化器学习更新模型参数。对数据集的获取，选择了中文 NER 引入词汇信息的开山之作，来自 ACL 2018 的论文 Chinese NER using Lattice LSTM 中从新浪财经收集的数据。标注集采用 BIOES 的格式（B 表示实体开头，E 表示实体结尾，I 表示在实体内部，LOC、PER 等表示具体的实体，O 表示非实体），句子与句子之间用一个空行隔开。具体如下：

```
新 B-ORG
华 M-ORG
社 E-ORG
华 B-GPE
盛 M-GPE
顿 E-GPE
4 O
月 O
2 O
8 O
日 O
```

```
电 O
（ O
记 O
者 O
翟 B-PER
景 M-PER
升 E-PER
） O
```

而在训练模型之前需要一些预处理。首先需要对数据集进行处理构建词表，建立字到数的映射以及分类到数的映射，除了训练集中的字，还需要添加’<PAD>’和’<UNK>’表示填充符（用于填充句子使一批样本长度相同进行训练）和未知字符（测试时出现而训练集中没有的字）。然后建立 MyDataset 类（继承 torch.utils.data.Dataset）实现自己的数据集，其功能是调用 Dataloader 封装时，可以指定参数 collate_fn 为自定义函数 pro_batch_data，实现多个句子长度对齐（填充’<PAD>‘）并转化为 GPU 上的张量。

```
def pro_batch_data(self,batch_datas):
    global device
    lenth=len(batch_datas[0][0])
    datas,tags=[],[]
    for data,tag in batch_datas:
        datas.append(data)
        tags.append(tag)

    datas = [i + [self.word2index["<PAD>"]] * (lenth - len(i)) for i in datas]
    tags = [i + [self.tag2index["<PAD>"]] * (lenth - len(i)) for i in tags]

    return
torch.tensor(datas,dtype=torch.int64,device=device),torch.tensor(tags,dtype=torch.int64,device=device)
```

四、系统主要模块流程

本项目的程序文件如下图所示：

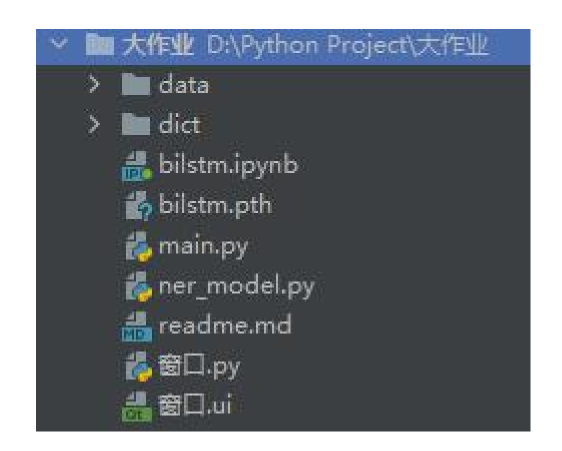


图 1 项目结构

文件主要包含以下几个部分，文件名与相对应功能解释如下：

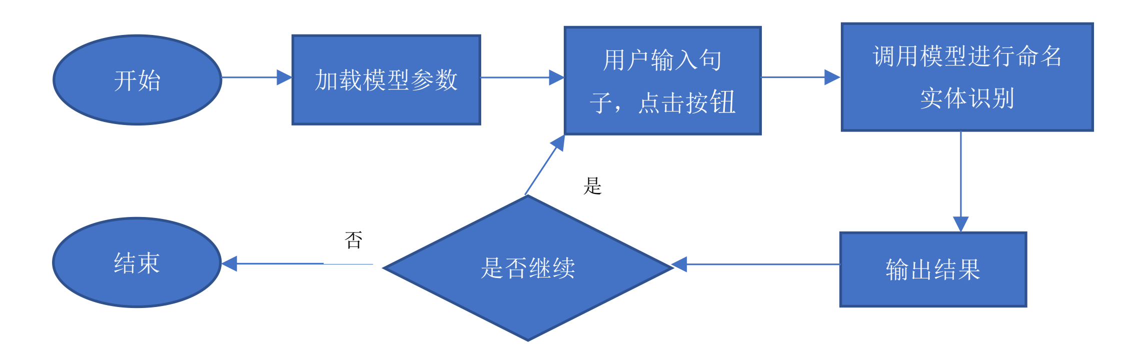
main.py	窗口主程序
窗口.ui	窗口原始 ui 文件
窗口.py	窗口对应 py 文件
ner_model.py	模型
bilstm.ipynb	训练模型使用的 jupyter 文件
bilstm.pth	模型参数
data	数据集
dict	词表

最终展示采用窗口形式，相应功能通过 pyqt5 实现。点击确定按钮便可以调用相应 NER 函数，通过之前训练好的模型对句子中的字进行分类，实现命名实体识别。



图 2 命名实体识别可视化界面

系统使用流程图如下：



五、实验结果及分析

30 次训练平均损失变化如下：

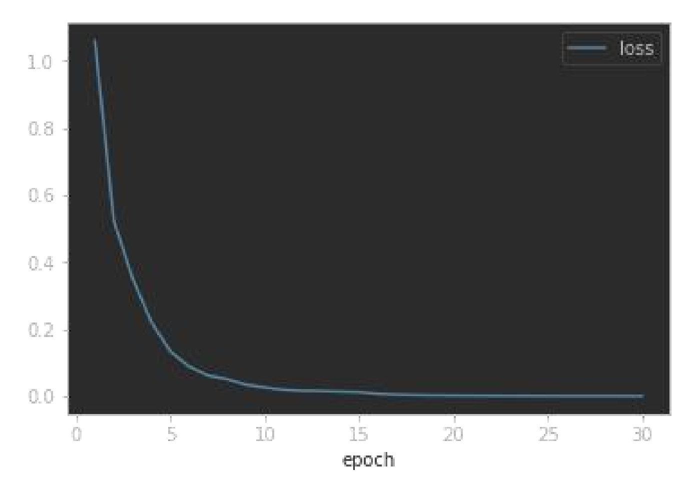


图 3 训练效果

训练过程较好，最后的在训练集上平均损失达到 0.003，训练集上 f1 值达到 0.99，而测试集 f1 值仅 0.94，存在一定过拟合现象，推测其原因：一是该数据集规模较小，训练集仅 50000 个字符；二是中文较为复杂，模型更难学习到数据间特征，导致缺乏良好的泛化性。

在经过多次随机输入一些语句的实验后，发现本系统对地名、机构名识别较为准确，但是人名识别准确率偏低，这与训练集实体大部分是地名、机构名有关。主要实验结果如下：

命名实体识别系统

请输入句子(中文)：

美国的华莱士，我和他谈笑风生。
香港记者跑得快。

确定

实体：

美 B-GPE
国 E-GPE
莱 M-GPE
士 E-GPE
香 B-GPE
港 E-GPE

命名实体识别系统

请输入句子(中文)：

美国第一任总统是华盛顿。
马龙是乒乓球运动员。

确定

实体：

美 B-GPE
国 E-GPE
华 S-GPE
盛 M-PER
顿 E-PER
马 B-GPE
龙 E-GPE

图 4 识别效果

命名实体识别系统

请输入句子(中文)：

中央银行宣布降息。
北京是中国的首都。

确定

实体：

中 B-ORG
央 M-ORG
银 M-ORG
行 E-ORG
北 B-GPE
京 E-GPE
中 B-GPE
国 E-GPE

命名实体识别系统

请输入句子(中文)：

平安银行在上海证券交易所上市。

确定

实体：

平 B-ORG
安 B-ORG
银 M-ORG
行 E-ORG
上 B-ORG
海 M-ORG
证 M-ORG
券 M-ORG
交 M-ORG
易 M-ORG
所 E-ORG

图 4 识别效果

参考文献：

[1] 李冬梅, 罗斯斯, 张小平, 等. 命名实体识别方法研究综述[J]. 计算机科学与探索, 2022, 16(9): 1954.

[2] Krupka G. SRA: Description of the SRA System as Used for MUC-6[C]//Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995. 1995.

[3] 张小衡, 王玲玲. 中文机构名称的识别与分析[J]. 中文信息学报, 1997, 11(4): 22-33.

[4] 王宁, 葛瑞芳, 苑春法, 等. 中文金融新闻中公司名的识别[J]. 中文信息学报, 2002, 16(2): 1-6.

[5] 罗凌, 杨志豪, 宋雅文, 等. 基于笔画 ELMo 和多任务学习的中文电子病历命名实体识别研究[J]. 计算机学报, 2020.

[6] Bikel D M, Schwartz R, Weischedel R M. An algorithm that learns what's in a name[J]. Machine learning, 1999, 34(1): 211-231.

[7] Zhou G D, Su J. Named entity recognition using an HMM-based chunk tagger[C]//Proceedings of the 40th annual meeting of the association for computational linguistics. 2002: 473-480.

[8] 俞鸿魁, 张华平, 刘群, 等. 基于层叠隐马尔可夫模型的中文命名实体识别[J]. 通信学报, 2006, 2.

[9] Borthwick A, Sterling J, Agichtein E, et al. Description of the MENE Named Entity System as used in MUC-7[C]//Proceedings of the Seventh Message Understanding Conference (MUC-7), Fairfax, Virginia, April 29-May 1, 1998. 1998.

[10] 张玥杰, 徐智婷, 薛向阳. 融合多特征的最大熵汉语命名实体识别模型[J]. 计算机研究与发展, 2008, 45(6): 1004-1010.

[11] Isozaki H, Kazawa H. Efficient support vector classifiers for named entity recognition[C]//COLING 2002: The 19th International Conference on Computational Linguistics. 2002.

[12] 陈霄, 刘慧, 陈玉泉. 基于支持向量机方法的中文组织机构名的识别[J]. 计算机应用研究, 2008, 25(2): 362-364.

[13] McCallum A, Li W. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons[J]. 2003.

[14] 燕杨, 文敦伟, 王云吉, 等. 基于层叠条件随机场的中文病历命名实体识别[J]. 吉林大学学报: 工学版, 2014 (6): 1843-1848.

[15] Han X, Kwoh C K, Kim J. Clustering based active learning for biomedical named entity recognition[C]//2016 International joint conference on neural networks (IJCNN). IEEE, 2016: 1253-1260.

[16] Li J, Sun A, Han J, et al. A survey on deep learning for named entity recognition[J]. IEEE Transactions on Knowledge and Data Engineering, 2020, 34(1): 50-70.

[17] Collobert R, Weston J, Bottou L, et al. Natural language processing (almost) from scratch[J]. Journal of machine learning research, 2011, 12(ARTICLE): 2493- 2537.

[18] Kong J, Zhang L, Jiang M, et al. Incorporating multi-level CNN and attention mechanism for Chinese clinical named entity recognition[J]. Journal of Biomedical Informatics, 2021, 116: 103737.

[19] Cetoli A, Bragaglia S, O'Harney A D, et al. Graph convolutional networks for named entity recognition[J]. arXiv preprint arXiv:1709.10053, 2017.

[20] Souza F, Nogueira R, Lotufo R. Portuguese named entity recognition using BERT-CRF[J]. arXiv preprint arXiv:1909.10649, 2019.

[21] Yang Z, Ma J, Chen H, et al. HiTRANS: A Hierarchical Transformer Network for Nested Named

Entity Recognition[C]//Findings of the Association for Computational Linguistics: EMNLP 2021. 2021: 124-132.